CLAIMS:

1. Method of extracting translations from translated texts, the method comprising the steps of:

accessing a first text in a first language;

accessing a second text in a second language, the second language being different from the first language, the second text being a translation of the first text;

dividing the first text and the second text each into a plurality of textual elements;

forming a sequence of pairs of text portions from said plurality of textual elements, each pair comprising a text portion of the first text and a text portion of the second text, each text portion comprising zero or more adjacent textual elements, each textual element of the first and the second text being comprised in a text portion of the sequence;

calculating a pair score of each pair in the sequence using the number of occurrences of each of a plurality of features in the text portions of the respective pair and using a plurality of weights, each weight being assigned to one feature of said plurality of features;

calculating an alignment score of the sequence using said pair scores, said alignment score indicating the translation quality of the sequence; and

optimizing said alignment score by systematically searching through the space of alternatives and combining optimal alignments for subsequences into optimal alignments for longer sequences.

2. The method of claim 1, wherein the dividing step includes a monolingual pre-processing step;

the monolingual pre-processing includes performing normalization of the textual elements, said normalization including lemmatization, case normalization, or truncation;

the monolingual pre-processing further includes counting the frequencies of the normalized textual elements that occur in the texts, and storing the frequencies;

the step of forming said sequence of pairs of text portions includes the steps of retrieving the stored frequencies and pairing text elements having at least similar frequencies; and

the method further comprises the step of reducing at least one weight assigned to a feature occurring in a text element pair if the difference between the frequencies of the paired textual elements exceeds a certain amount.

3. The method of claim 1, wherein said pair scores are calculated by taking, for each feature occurring in the pair, the minimum number of the numbers of occurrences of the respective feature in the paired text portions, taking the product of said minimum number and the weight assigned to the respective feature, and summing up all said products of all features; and wherein:

the alignment score is calculated by summing up all the pair scores; and

the alignment score is optimized by selecting the maximum alignment score.

4. The method of claim 1, wherein said plurality of features include lexical information.

5. The method of claim 1, wherein said plurality of features include document structure and formatting information.

6. The method of claim 1, wherein said plurality of features include any character within the text, the weights of such features being lower than the weights of other features.

7. The method of claim 1, further comprising the step of generating pairs of textual elements, each pair comprising a textual element of the first text and a textual element of the second text;

wherein the step of generating pairs of textual elements comprises the step of normalizing textual elements; and

wherein the normalizing step includes removing accents, inessential non-alphanumeric characters, or case normalization.

8. The method of claim 1, further comprising the step of generating pairs of textual elements, each pair comprising a textual element of the first text and a textual element of the second text;

wherein the step of generating pairs of textual elements comprises the step of accessing at least one bilingual resource.

9. The method of claim 1, wherein the first and second texts are provided in the form of a first and second document, the first and second languages being natural languages, and wherein the method is used for extracting sentence translations.

10. The method of claim 1, wherein the first and second texts are provided in the form of speech signals and a transcript thereof.

11. The method of claim 1, wherein the first and second texts are related DNA sequences.

12. The method of claim 1, wherein the forming, calculating and optimizing steps are performed in a dynamic programming process comprising the steps of:

accessing a set of nodes, each node being a pair of positions in the first and second texts, each node being annotated with a node score;

for each node, generating a set of successor nodes by applying a set of node transitions; and

for each successor node, calculating a node score using the node score of the node accessed for generating the successor nodes.

13. The method of claim 12, wherein said node score is the score of the best alignment that led to the respective node; and wherein:

each node has assigned a pointer to a predecessor node that took part in the best alignment that led to the respective node; and

the process further comprises the step of deleting each node which has no successor node that points to the node as its predecessor node.

14. The method of claim 12, wherein the process further comprises a pruning step of comparing the score of each successor node with the scores of competing nodes spanning a similar part of the first and second texts, and deleting those successor nodes having scores being considerably worse than the scores of the competing nodes.

15. The method of claim 14, further comprising the steps of:

estimating the number of matches that can be achieved in the alignment of the remaining parts of the texts; and

using said estimate in comparing the competing nodes.

16. The method of claim 14, further comprising the steps of

computing an approximate alignment before performing the forming, calculating and optimizing steps; and

using said approximate alignment in estimating the number of matches.

17. The method of claim 14, wherein the step of estimating the number of matches includes the step of accessing an index for determining for each feature occurrence where in the respective text the feature occurs.

18. The method of claim 14, further comprising the step of performing a backward run of the Hunt/Szymanski algorithm and recording the intermediate results sequentially in a stack such that they can be retrieved in reverse order.

09673950 . 121500

19. A computer readable storage medium storing instructions for performing a method comprising the steps of:

accessing a first text in a first language;

accessing a second text in a second language, the second language being different from the first language, the second text being a translation of the first text;

dividing the first text and the second text each into a plurality of textual elements;

forming a sequence of pairs of text portions from said plurality of textual elements, each pair comprising a text portion of the first text and a text portion of the second text, each text portion comprising zero or more adjacent textual elements, each textual element of the first and the second text being comprised in a text portion of the sequence;

calculating a pair score of each pair in the sequence using the number of occurrences of each of a plurality of features in the text portions of the respective pair and using a plurality of weights, each weight being assigned to one feature of said plurality of features;

calculating an alignment score of the sequence using said pair scores, said alignment score indicating the translation quality of the sequence; and

optimizing said alignment score by systematically searching through the space of alternatives and combining optimal alignments for subsequences into optimal alignments for longer sequences.

20. A system for extracting translations from translated texts, comprising:

a pre-processor for accessing a first text in a first language, accessing a second text in a second language, the second language being different from the first language, the second text being a translation of the first text, and dividing the first and the second text each into a plurality of textual elements; and

a processor for forming a sequence of pairs of text portions from said pluralities of textual elements, each pair comprising a text portion of the first text and a text portion of the second text, each text portion comprising zero or more adjacent textual elements, each textual element of the first and the second text being comprised in a text portion of the sequence, the processor being further arranged for calculating a pair score of each pair in the sequence using the number of occurrences of each of a plurality of features in the text portions of the respective pair and using a plurality of weights, each weight being assigned to one feature of said plurality of features, calculating an alignment score of the sequence using said pair scores, said alignment score indicating the translation quality of the sequence, and optimizing said alignment score by repeating said forming and calculating steps.